IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-3 E-ISSN: 2347-2693

Prediction Model for Diabetes Mellitus Using Machine Learning Techniques

N.A. Farooqui[1*], Ritika[2], A. Tyagi[3]

[1*]Computer Applications, DIT University, Dehradun, India
[2] Computer Applications, DIT University, Dehradun, India
[3] Computer Applications, DIT University, Dehradun, India

[*]*Corresponding Author: farooqui_mca@yahoo.com, Tel.: +91-94543-24373*

Available online at: www.ijcseonline.org

Received: 24/Feb//2018, Revised: 03/Mar2018, Accepted: 23/Mar/2018, Published: 30/Mar/2018

***Abstract*—** In today's world diabetes is the major health challenges in India. It is a group of a syndrome that results in too much sugar in the blood. It is a protracted condition that affects the way the body mechanizes the blood sugar. Prevention and prediction of diabetes mellitus is increasingly gaining interest in medical sciences. The aim is how to predict at an early stage of diabetes using different machine learning techniques. In this paper basically, we use well-known classification that are Decision tree, K- Nearest Neighbors, Support Vector Machine, and Random forest. These classification techniques used with Pima Indians diabetes dataset. Therefore, we predict diabetes at different stage and analyze the performance of different classification techniques. We Also proposed a conceptual model for the prediction of diabetes mellitus using different machine learning techniques. In this paper we also compare the accuracy of the different machine learning techniques to finding the diabetes mellitus at early stage.

***Keywords*—** Diabetes; Decision Tree, K-Nearest Neighbors, Machine Learning, Random Forest, Support Vector Machine.

I. INTRODUCTION

According to world health organization (WHO) and International Diabetes Federation (IDF) the number of diabetes patients increases in India in last 10 years. Diabetes is one of the common disease and it is broadly classified into three categories namely Type1 (Juvenile diabetes), Type2(insulin dependent diabetes), gestational diabetes. The annual report 2016 of International diabetes federation, reports that diabetes is the top two chronic diseases in India. Last year on November the regional meeting of the International diabetes federation South East Asia (IDF SEA) region was held in Hyderabad that report's shows that 1 of 12 Indian had diabetes and more than 65.1 million people suffers from diabetes [1]. The report also indicates that the number of diabetes patients is increased every year. According to data published by the global burden of disease (GDB) in 2016, 347,000 people died of diabetes, which caused 3.3% of all deaths that year, with an annual increase of 2.7% from 1990. Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion [2]. Insulin deficiency increases the glucose levels in the blood and reduced the metabolism of carbohydrates, fat and proteins. DM is the most common endocrine disorders, affecting more than 415million person in the whole world. DM evolution is strongly linked to several complications, mainly due to chronic hyperglycemia. DM covers a wide range of heterogeneous pathophysiological conditions. The most common complications are divided into micro- and macro-vascular disorders, including diabetic nephropathy, retinopathy, neuropathy, diabetic coma and cardiovascular disease. Due to high DM death rate and indisposition as well as related disorders, prevention and treatment attracts broad and significant interest. Insulin is the main treatment for Type1, even though in certain cases insulin is also provided to Type2 diabetic patients, when hyperglycemia cannot be controlled through diet, weight loss, exercise and oral medication. The most common anti-diabetic agents include sulfonylurea, metformin, alpha glucosidase inhibitor, peptide analogy, non-sulfonylurea, secretagogues, etc. [3].

Most of the present anti-diabetic agents, exhibit numerous side-effects. In addition, insulin therapy is related to weight gain and hypoglycemics events. Hence, anti-diabetic drug design and discovery is of great concern and concurrently a research challenge [4,5,6,7]. During the last decades there should be advance research in DM provides a huge and important knowledge, on the a) etiopathology (genetic or environmental factors and cellular mechanisms), b) treatment, and c) screening of the disease, there is still much to be discovered, ragged, clarified and described. Through such processes diagnosis, predictive evaluation of

appropriate treatment and clinical administration will handle the diabetes. Thus, such an effort, will reduced the mortality rate and establish the safe treatment. Thus, Machine learning is the key factor to take the decision by doctor and very helpful for the diagnosis and appropriate decision making in drug administration.

Machine learning is prominent technique in medical sciences. This is promising approach that improves sensitivity and specificity of disease detection and diagnosis by K-Nearest Neighbors, Decision tree, Support Vector Machine and Random forest [8]. In this paper, the disease diagnosis and decision making is obtained from the Pima Indian diabetic database [9]. The objective of this study is to evaluate the performance of machine learning techniques to classify patients from diabetic mellitus using three different ordinal adult groups namely (i) young adults (ii) middle age adults (iii) adults older than 55.

II. LITERATURE REVIEW

The Most popular system LDA–MWSVM proposed by Calisir and Dogantekin for diagnosis of the diabetic patients [10]. The system performs feature extraction and reduction using the Linear Discriminant Analysis (LDA) method, followed by classification using the Morlet Wavelet Support Vector Machine (MWSVM) classifier. Gangji and Abadeh [11] proposed a classification system that is based on an Ant Colon, which have set of fuzzy rules named FCSANTMINER for diagnosis of diabetes. Multivariate regression problem is solved by the Support Vector Regression (SVR), that is useful in the prediction of the glucose level at different stages of diabetes mellitus [12]. Agarwal et al. [13] creates the phenotype machine learning methods for the diabetic mellitus by utilizing the semi-automatically labelled training sets.

In [14], authors proposed a fuzzy ontology-based, Case-based reasoning(CBR) framework used for the diagnosis of diabetes. A hybrid model was proposed by Chin-Yuan Fan et al [15], by integrating a case-based data clustering method and a fuzzy decision tree to classify the liver disorder and breast cancer datasets. Nihat et al [16], proposed modified K-Means for removing noisy data and SVM for classification of the reduced datasets.

Yuan et al [17], has modelled an objective function which is based on the leave-one-out cross validation, and the SVM parameters are optimized by using GA and PSO (particle swarm optimization). Patil et al [18], proposed a hybrid K-Means followed by Naive Bayes and SVM classification, in which SVM achieved high accuracy percentage. Aishwarya et al [19], proposed a medical decision support system based on genetic algorithm to extract features and Least Square SVM for diagnosis of diabetes.

III. METHODS AND MATERIAL

The disease prediction and diagnosis use the different algorithms and approaches which can be applied by the traditional machine learning algorithms, ensemble learning approaches and association rule learning to achieve the best classification accuracy [14,20]. The extensive afford made to identify machine learning techniques to diabetic research. Two databases searched: the one used in medical sciences and other in computer sciences. There is a close relationship between machine learning and datamining thus, some time machine learning techniques are called datamining techniques [21].

In this study we obtain a dataset from Pima Indian diabetes database that have several features that are responsible for diagnosis and prediction. We proposed the Prediction model for the diabetes mellitus that is shown in the fig1. In this model take the training data from the Pima Indian diabetes dataset which are noise free and apply the machine learning techniques to evaluate the training data then create a model, which is implemented by the different machine learning prediction classifier and take the decision based on the input data. There is limitation for the input data is that it is a noise free. In this model we supposed that data is not missing and semi structured

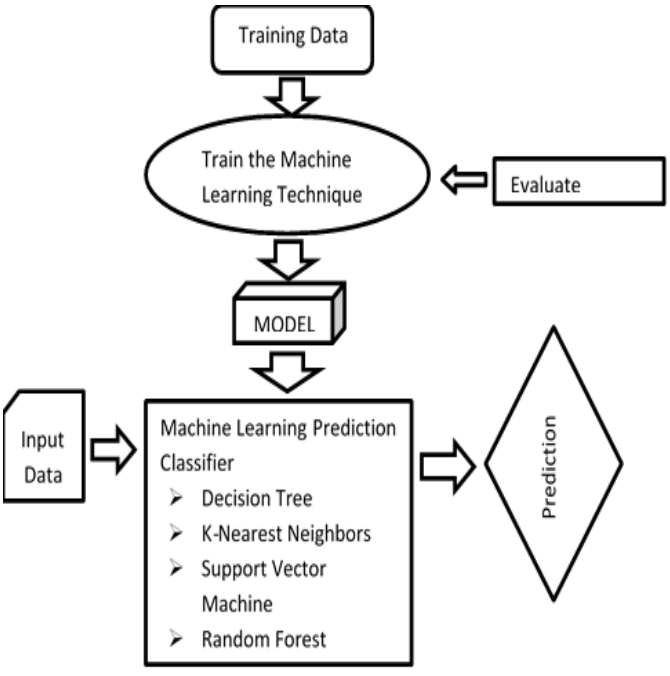

Figure1: Conceptual Prediction Model

A. Input Data Processing

An initial data is collected from Pima Indians diabetes dataset. It will be used to comparative analysis of different machine learning techniques [8]. There are some variables are needed for screening of the diabetic patient prediction.

Table I. Input and Output variables

S. No.	Variables	Description	Value
1	AGE	Age(Years)	Numeric
2	WEIGHT	Weight(Kg)	Numeric
3	BMI	Body Mass Index(Kg/m^2)	Numeric
4	BPH	Systolic Blood Pressure(mmHg)	Numeric
5	BPL	Diastolic Blood Pressure(mmHg)	Numeric
6	DM_FAMILY	History of Diabetes in Family	Yes/No/Unknown
7	HT_FAMILY	History of Hypertension in Family	Yes/No/Unknown
8	SMOKE	Smoking Habit	Yes/No/Occasionally/Unknown
9	SEX	Sex	Male/Female
10	CLASS	Diabetic and Nondiabetic Group	Two Groups

B. Predictive Models

The following Machine learning techniques are used for comparative analysis for the predictive model of diabetes.

a. Decision tree.
b. K-Nearest Neighbours.
c. Support Vector Machine.
d. Random forest.

a. Decision tree

A Decision Tree is a classifier using the classification regression trees(CART) algorithm that is capable of handling both classification and regression unlike simple decision tree algorithm. It does not have a computational set of rules. Generally, we construct using 'divide and conquer' strategy.

Algorithm to generate decision tree

Input:

a. Set of input data are training samples.
b. Set of attributes from input samples.
c. Splitting the attributes by best partitioning criteria.
Output: A decision tree.

Method:

1) Create a node N.
2) If samples S are in same class C. Then
3) Return N with labelled class C.
4) If attributes list is empty, then
5) Return N with most of class in S.
6) The best information gain of an attribute. Select which has the highest information gain.
7) It creates a decision node with that attribute.
8) Repeatedly apply the process and nodes are added to its child node

b. K-Nearest Neighbour

KNN is a classification algorithm. Where we measure the distance between the objects by the distance function. The Euclidean distance between two points a and b using two planes X and Y are given by the equation.

$$\text{Euclidean Dist. ((X, a), (Y, b))} = \sqrt{\sum_{i=1}^{k}(b-a)^2}$$

Table II. Euclidean Values

K	Euclidean values
4	**1.191145**
5	**1.092156**
6	**1.016752**

c. Support Vector Machine

A Support Vector Machine is a supervised classification algorithm that has been extensively and successfully used for text classification task. That supports the regression and classification tasks and can be handled with multiple variables. The main purpose of this approach is the prediction of the membership of the class to categorize into different classes using hyperplane.

d. Random forest

Random Forest is the one of the Classifier which is used for Classifications problems. Random Forest is ensemble classifier made using many decision trees where ensemble means that uses multiple machine learning algorithm to obtain the predictive performance It is better than other for the prediction of diabetes mellitus. This algorithm is as follow:

1 Draw N-Tree bootstrap sample from the input data.
2 For each of the bootstrap sample, grow an unpruned regression by splitting the node from all predictor nodes. The predictors choose the best split from input variables. (This thought is called bagging)
3 Predict new data by aggregating the prediction of N-Trees

IV. MEASUREMENT AND COMPARATIVE ANALYSIS

To compare the performance of different techniques by measuring the accuracy. Suppose TP, FP, TN and FN is the number of true positive, false positives, true negatives and false negatives respectively [22,23]. Therefore, accuracy is defined as:

Accuracy= (TP+TN)/ (TP+TN+FP+FN)

Table3. Comparison Result of Different Classification Techniques

S. No.	Classification Techniques		Accuracy (%)
1	Decision Tree		84.05
2	K-Nearest Neighbours	K=4	86.69
		K=5	82.55
		K=6	80.01
3	Support Vector Machine		80.85
4	Random Forest		96.89

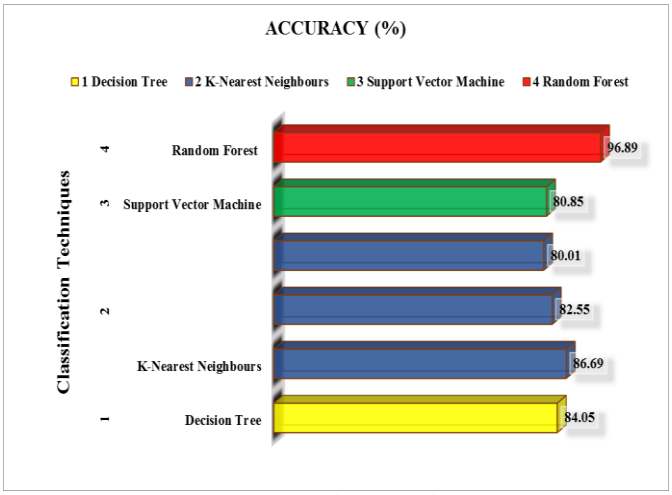

Figure2. Graphical representation of the Classification techniques (DT, KNN, SVM, RF)

V. CONCLUSION

In this paper, we have used different machine learning techniques viz. Decision tree, K-nearest Neighbours, Random Forest and Support Vector Machine and predict the performance of different classification techniques. It further proposes the conceptual prediction model that includes the different machine learning classifiers. Thus, by comparing the accuracy of different machine learning techniques; it concludes that Random Forest performance is better than other Classification Techniques. This analysis can be carried out on different models for other diseases with suitable datasets in future.

ACKNOWLEDGMENT

The author thanks the DIT University, Dehradun for providing the research grant to support this research work. The corresponding author wishes to thanks Prof K.K. Raina and Prof S.K. Gupta for the great cooperation and motivation for this research.

REFERENCES

[1] International Diabetes federation. Diabetic Atlas Fifth Edition 2011.

[2] American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care 2009;32(Suppl. 1): S62–7.

[3] Krentz AJ, Bailey CJ. Oral antidiabetic agents: current role in type 2 diabetes mellitus. Drugs 2005;65(3):385–411.

[4] Tsave O, Halevas E, Yavropoulou MP, Kosmidis Papadimitriou A, Yovos JG, Hatzidimitriou A, et al. Structure-specific adipogenic capacity of novel, welldefined ternary Zn(II)-Schiff base materials. Biomolecular correlations in zincinduced differentiation of 3T3-L1 pre-adipocytes to adipocytes. J Inorg Biochem Nov 2015; 152:123–37.

[5] Halevas E, Tsave O, Yavropoulou MP, Hatzidimitriou A, Yovos JG, Psycharis V, et al. Design, synthesis and characterization of novel binary V(V)-Schiff base materials linked with insulin-mimetic vanadium-induced differentiation of 3T3-L1 fibroblasts to adipocytes. Structure–function correlations at the molecular level. J Inorg Biochem Jun 2015; 147:99–115.

[6] Tsave O, Yavropoulou MP, Kafantari M, Gabriel C, Yovos JG, Salifoglou A. The adipogenic potential of Cr(III). A molecular approach exemplifying metalinduced enhancement of insulin mimesis in diabetes mellitus II. J Inorg Biochem Oct 2016; 163:323–31.

[7] Sakurai H, Kojima Y, Yoshikawa Y, Kawabe K, Yasui H. Antidiabetic vanadium(IV) and zinc(II) complexes review article coordination. Chem Rev March 2002; 226(1–2):187–98.

[8] Nongyao Nai-arun, Rungruttikarn Moungmai(2015)Comparison of classifiers for the risk of diabetes ELSEVIER Procedia Computer Science 69 (2015) 132-142.

[9] Pima Indian diabetes datasets from UCI Repository.

[10] Çalisir D, Dogantekin E. An automatic diabetes diagnosis system based on LDA Wavelet Support Vector Machine Classifier. Expert Syst Appl 2011;38(7):8311–5.

[11] Ganji MF, Abadeh MS. A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis. Expert Syst Appl 2011;38(12):14650–9.

[12] Georga EI, Protopappas VC, Ardigò D, Marina M, Zavaroni I, Polyzos D, et al. Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. IEEE J Biomed Health Inform 2013; (1):71–81.

[13] Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc May 12, 2016.

[14] El-Sappagh S, Elmogy M, Riad AM. A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis. Artif Intell Med Nov 2015;65(3):179208.

[15] Chin-Yuan Fan, Pei-Chann Chang, Jyun-Jie Lin, J.C. Hsieh. A Hybrid model combining Case-based reasoning and Fuzzy Decision Tree for Medical Data Classification. Applied Soft Computing; 2011; 11(1); pp.632–644.

[16] Nihat Yilmaz, Onur Inan, Mustafa Serter Uzer. A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases. Springer: Transaction Processing Systems:J Med Syst; April 2014;38:48.

[17] Yuan R, Guangchen B. Determination of Optimal SVM Parameters by Using Genetic Algorithm/Particle Swarm Optimization. Journal of Computers; 2010; No.5;1160-1169.

[18] Patil B M, Joshi R C, Toshniwal D. Impact of K-Means on the performance of classifiers for labeled data. Comm. Com. Inf. Sc.94;2010; 423-434.

[19] Aishwarya S, Anto S. A Medical Decision Support System based on Genetic Algorithm and Least Square Support Vector Machine for Diabetes Disease Diagnosis. International Journal of Engineering Sciences & Research Technology; April2014; 3(4).

[20] OhW, KimE, CastroMR, Caraballo PJ, Kumar V, Steinbach MS, et al. Type 2 diabetes mellitus trajectories and associated risks. Big Data Mar 1 2016;4(1):25–30.

[21] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda Machine learning and datamining methods in diabetes research ELSEVIER Computational and Structural Biotechnology journal 15(2017) 104-116.

[22] J. Pradeep Kandhasamy, S. Balamurali(2015) Performance analysis of classifier models to predict diabetes mellitus ELSEVIER Procedia Computer Science 47 (2015) 45-5.

[23] N. Radha and S. Ramya "Performance Analysis of Machine Learning Algorithms for Predicting Chronic Kidney Disease", International Journal of Computer Sciences and Engineering Vol.-3, Issue -8, pp(72-76) Aug 2015 E-ISSN: 2347-2693.

Authors Profile

Mr. N A Farooqui recieved Bachelor of Science(Hons) from Aligarh Muslim University, Aligarh in 2005 and Master of Computer Application from Integral University, Lucknow in year 2010. He is currently pursuing Ph.D. and currently working as Research Associate in Department of Computer Applications, DIT University, Dehradun since 2017. He is a member of International Association of Engineers (IAENG) since 2017, ACM since 2011. He has published more than 10 research papers in reputed international journals including conferences proceeding and attended Workshops and FDP's during the teaching. His main research work focuses on Machine Learning, Data Mining, Artificial Intelligence based education. He has 7 years of teaching experience and guided many projects during teaching.

Dr.Ritika (CSI Life time membership No: 00128568) is an Associate Professor in the department of Computer Science at DIT University. She received her Ph.D. degree in Computer Science from Gurukul Kangri University, Haridwar, M.Tech degree in Computer Science and Engineering from Uttarakhand Technical University, Dehradun. She specializes in core areas of computer science and holds experience of more than 16 years. She is an innovative person with deep knowledge of Advance Networking, Mobile Computing, Data warehousing and mining etc. She has also published many research papers in various National and International Journals.

Ms Ankita Tyagi received Bachelor of Computer Application (BCA) from GGSIP Univesity , Delhi in 2012 and Master of Computer Application from GGSIP university, Delhi in 2016. She is a memberof International Assosciation of engineers (IAENG) since 2018. Her main research work focusses on machine learning, Bigdata, Bionformatcs and Artificial Intelligence.

